DOCUMENT RESUME

ED 475 829 TM 034 853

AUTHOR	Luecht, Richard M.
TITLE	Applications of Multidimensional Diagnostic Scoring for Certification and Licensure Tests.
PUB DATE	2003-04-00
NOTE	34p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Certification; Diagnostic Tests; *Licensing Examinations (Professions); *Reports; *Scores; *Scoring; Simulation
IDENTIFIERS	*Multidimensionality (Tests)

ABSTRACT

This paper discusses two topics related to diagnostic score reporting for credentialing examinations. The first deals with various ways to compute subscores for credentialing examinations. The second addresses some pertinent factors to consider when presenting diagnostic results. To illustrate these issues, a sample set of subscores is used. This set was derived from a certification test that provides pass/fail decisions on multiple sections. There are a number of ways to compute diagnostic subscores for competency areas; the paper discusses four approaches. A simulation study using these approaches shows the complexity of choosing a scoring model for multidimensional subscore reporting. The decision to use a given method to compute diagnostic scores should blend technical sophistication with operational needs. There is very little research literature on presenting scores, but there are a number of techniques from which to choose, including score tables, profile plots, and narrative text. Producing high quality score reports is feasible even for relatively small testing programs. (Contains 5 figures, 3 tables, and 22 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 475 829

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
R. M. Luecht
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

Applications of Multidimensional Diagnostic Scoring for Certification and Licensure Tests

Richard M. Luecht

University of North Carolina at Greensboro

April 2003

TM034853

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL

BEST COPY AVAILABLE

Certification and licensure tests provide a means of credentialing individuals for various occupations and professions. Federal, state, or local governments may mandate licensure testing in various professions to verify that the person granted the license has a sufficient degree of the required knowledge and skills to effectively and safely perform their occupational responsibilities. Examples include physicians, nurses, and airline pilots. Similarly, certification tests are used by many non-government-regulated professions to credential individuals as having attained some designated level of mastery on a set of competencies. The competencies may represent broad levels of knowledge and skills or specific skills (e.g., successfully using tax preparation software). Some certification examinations assess the minimum competencies required by entry-level professionals. Other examinations assess "accomplished" or "excellent" sustained performance that [usually] may only be attained by the most experienced and highly skilled professionals (e.g., master teachers). In the end, a candidate either passes or fails at some prescribed level of competence.

Some argue that candidates ought to be satisfied just to know whether they passed or failed the credentialing test. In fact, a recurring debate among professional examination policy groups is whether or not scores should even be reported. So, why do we need scores on credentialing examinations? Sometimes, a picture is better than words. Figure 1 provides a whimsical view of a score report without any scores. It takes only a small amount of imagination to notice

the similarities to sweepstakes announcements printed inside candy wrappers or under soda water bottle caps..."*[Pay your money and] please try again!*".

[INSERT FIGURE 1 ABOUT HERE]

There are, of course, a number of positive reasons to report scores, especially diagnostic scores. For example, failing candidates almost always want to know how they did on particular parts of the test—especially where they did most poorly—to help them study for a retest. In addition, faculty at colleges or universities that have professional training programs sometimes use individual or aggregate test results from their students to gauge the success of their curricula and make necessary modifications[1].

At the same time, score reporting for certification and licensure tests can be technically complicated and costly. For example, small candidate samples encountered in some professions make it difficult to estimate stable item and score statistics, especially for some of the "information hungry" IRT models used to calibrate and equate these tests. Another complication is that many professions are comprised of homogeneous, highly proficient examinee populations that provide only limited "person information" for item calibration purposes. Finally, organizations need to consider the costs of item development to support score reporting. That is, there are incremental costs associated with building and maintaining item banks that would otherwise be designed strictly

---

[1]Some testing organizations have arrangements with professional schools to report anonymous aggregate results.

to maximize score precision in the region of the pass/fail cut score. The latter point is an especially germane issue for diagnostic score reporting. Item banks that are optimally designed for making pass/fail decisions tend to be rather homogeneous in terms of item difficulty and may lack adequate test information for score reporting purposes away from the cut score. Adaptive strategies using multistage tests can help (Luecht and Nungester, 1998), but the item banks, themselves, can severely limit the degree of adaptation possible (Xing & Hambleton, 2001; Luecht, 2003).

Is there an obligation to report diagnostic scores for credentialing examinations? It depends. Unless required by state or local governmental professional licensing authorities or mandated by the courts there is no legal obligation for a testing organization to report diagnostic information on a credentialing examination. In fact, the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) seem to avoid taking any strong position about obligations of certification and licensure testing authorities to report diagnostic scores[2]. Nonetheless, many organizations—either in response to pressure from influential constituents, or, out of a sense of ethical or moral obligation—provide some level of diagnostic feedback information to candidates.

Given this general context for credentialing examinations, this paper discusses two topics related to diagnostic score reporting for credentialing

---

[2] If subscores are reported, the *Standards* do recommend reporting reliability and validity evidence.

examinations. The first, rather technical topic deals with various ways to compute subscores for credentialing examinations. The second topic addresses some pertinent factors to consider when presenting diagnostic results. Both topics are equally relevant.

The phrase "diagnostic" implies useful feedback information for detecting and evaluating an examinee's strengths and weaknesses. However, there are two important aspects of "information": (1) the *presentation form* of data for some explicit informational purpose—here to provide useful feedback to examinees and (2) the nature of the *data* to be presented. In testing contexts, all information provided to candidates is expected to be reliable and valid. We certainly do not want examinees pursuing improper remedial actions based up faulty diagnostic information.

What is sometimes overlooked is that both the data and its presentation need to meet acceptable quality standards. For example, producing an elegant line graphic to display a multivariate profile of some number of unreliable subscores is clearly not a recommended psychometric practice. Conversely, presenting highly reliable and valid subscores in a poorly designed graphic display that distorts the meaning of the data is not recommended. Effective presentation of data is a matter of displaying perceptible information that accurately, clearly and efficiently conveys a particular meaning. Graphics, tables, and verbal descriptions can all present the same data, but each may reveal different information that conveys unique meanings (Tufte, 2001; Wainer &

Thissen, 1981). Ultimately, accurate and valid performance-based data becomes diagnostic information when presented in a form that facilitates appropriate understanding by the candidates (or other authorized users of the information).

A Data Source for Exploration and Discussion

It is helpful to have a sample data set of subscores to illustrate some issues discussed further on in this paper. This example set of subscores was derived from a certification test that provides pass/fail decisions on multiple sections; although the data set is based on only one section that covers four broad professional competency areas. These are simply called Competency Areas #1, #2, #3, and #4.

Figure 2 shows the relative test information curves for each of the four competency areas (the dashed curves). The vertical solid line is the approximate cut score on this test. Each of the test information curves therefore peaks near to that cut score. An assumed cumulative normal ability distribution is also displayed (the solid curve). The test is obviously most informative for examinees nearer to the cut score and rapidly becomes less informative toward the tails of the proficiency distribution.

[INSERT FIGURE 2 ABOUT HERE]

This examination section had seventy-four items calibrated from a recent large-scale administration of the test, using the three parameter logistic (3PL) IRT model. The IRT item statistics are summarized in Table 1, by competency area.

There are noticeable differences in the average discrimination parameters as well as in the item difficulties. The differences in item counts are also relevant insofar as affecting the reliability of the individual subscores.

[INSERT TABLE 1 ABOUT HERE]

The 3PL item statistics from the operational calibration were treated as known parameters and integrated into a four-factor, oblique simple-structure multidimensional IRT (MIRT) model. The MIRT model was to generate response data for 2000 simulated examinees. Each item was forced to load on only one of the four competency traits. A vector of four trait scores was sampled from multivariate normal distribution for each of the simulated examinees. Pair-wise correlations of 0.50 were induced between the four traits, producing the oblique simple structure. Finally, using a well-known IRT-based response generating mechanism involving uniform probabilities, a 2000 × 74 matrix of dichotomous responses was produced. Each of the 2000 simulated examinees had responses to all seventy-four items.

Computing Diagnostic Scores for Credentialing Examinations

As noted earlier, credentialing examinations are mastery tests; the candidates either pass or fail. These types of tests are usually not meant to measure a wide range of achievement or aptitudes. Nonetheless, most credentialing examinations are constructed on the basis of an evidence-based "competency model" that follows from a formal practice or job analysis.

If we consider the competencies to be distinct traits then most credentialing tests are, to some degree, multidimensional. That is, although a single, total-test score may ultimately be used for making the pass/fail decisions, there is an implicit assumption that the test is mixture of multidimensional traits. Going a step further, it should be possible to define a structure that appropriately represents the professional knowledge and skill competency traits (factors) underlying each of the competency-based subscores to be reported (Luecht, 1996). By treating the diagnostic subscore space as multidimensional, we acknowledge that each competency contributes some amount of unique variance to the overall trait. Furthermore, by explicitly modeling the multidimensional structure of the data, we can capitalize on shared information among the appropriate item responses to ideally improve the reliability of the estimated subscores. In a certification or licensure setting, where the statistical test information available for score reporting is sparse in some regions of the scale(s), collateral information, gleaned from the covariances among the response variables, could help improve the reliability of otherwise marginally reliable scores.

Still, dealing with multidimensional IRT models in operational settings is not always straightforward, despite many recent advances in IRT modeling and estimation algorithms. Some of the technical complications include making subtle choices between multidimensional models, choosing among different estimators (e.g., maximum likelihood, Bayes mean and mode estimators),

confirming the number of dimensions and structural invariance of the trait space, dealing with empirical factor identification issues, and rotational indeterminacies. There are also issues related to the theoretical or substantive specification and interpretation of the multidimensional space (e.g., simple content-based factors versus cognitive scientific perspectives on task demands and multidimensionality).

Methods of Computing Diagnostic Subscores

The challenge is to report reasonable diagnostic subscores for each of the four competency areas. There are a number of ways to proceed. Here, I considered four approaches: (i) using standardized number correct scores within each of the competency areas, denoted ZX; (ii) computing Bayes mean or *expected a posteriori* (EAP) scores[3] based on a unidimensional total-test calibration of the items, denoted UIRT(T), (iii) computing Bayes mode or *maximum a posteriori* (MAP) scores based on separate unidimensional calibrations of items for the separate competency area, denoted as UIRT(S); and (iv) MAP scores based on a multidimensional calibration of the entire test, with one factor representing each competency area (MIRT).

The UIRT(T) approach was implemented by calibrating all of the items on the test using a unidimensional IRT model like the three-parameter logistic model (normal ogive approximation):

---

[3] As a processing convenience, a separate scoring program, SCORE3PL (Luecht, 1999), was used to compute the EAP scores for the four competency areas, based on the total-test calibrations. SCORE3PL does not compute MAP scores. All MAP scores were computed directly by the calibration software packages, BILOGMG-3 and TESTFACT4.

$$P(u_i|\theta) = c_i + \frac{1 - c_i}{1 + exp[-1.7a_i(\theta - b_i)]} \quad (1)$$

In Equation 1, u_i is a binary item response; a_i, b_i, and c_i, are the usual item parameters (see, for example, Hambleton and Swaminathan, 1985). Scoring was performed as a separate step, using the subsets of item parameter estimates and the corresponding examinee responses for each of the four competency areas. This method was selected because it offers the advantage of requiring only a single calibration of the items. On the downside, this UIRT(T) method loses some of the unique variance associated with the individual factors and can induce an upward correlation bias between the subscores.

In contrast, the UIRT(S) approach treats each competency area as a separate subtest. Separate calibrations are needed for each subtest, again using a unidimensional model like the 3PL model. This approach was selected because it has two apparent advantages: (a) the independent calibrations allow the trait composites for that the individual competency metrics to diverge from each other, as necessary, and (b) those metrics can be independently equated or linked to provide comparisons across time.

The fourth approach investigated was to use a multidimensional item response theory (MIRT) model to calibrate all of the items on the test. The normal ogive approximation to the three-parameter logistic model was used (Bock et al, 2003):

$$P(u_i|\theta) = c_i + \frac{1-c_i}{1+exp\left[-1.7\left(\mathbf{a}_i^T\theta + d_i\right)\right]} = c_i + \frac{1-c_i}{1+exp\left[-1.7\left(a_{i1}\theta_1 + a_{i2}\theta_2 + \cdots + d_i\right)\right]} \quad (2)$$

where $\mathbf{a}_i = (a_{i1}, a_{i2,\ldots}, a_{im})$ is a vector of item coefficients (loadings), d_i, is the item difficulty parameter, and c_i is a pseudo-guessing parameter. For a single examinee, the vector of proficiencies, θ, has the same order as $\mathbf{a}_i$. It is important to realize that there are several variants of MIRT models, including the oblique simple structure model (Thurstone, 1947), the common-factor model, and the bi-factor model (Gibbons and Hedecker, 1992). Unfortunately, there is seems to be a lack of solid operational research about these variants and only limited available software for carrying out the calibrations. Furthermore, as demonstrated below, the results of a MIRT analyses are not always easily comparable with unidimensional methods, except in terms of model fit.

The 2000 × 74 matrix of dichotomous item responses was used to compute four scores for each simulated examinee: ZX, UIRT(T), UIRT(S), and MIRT. BILOG-MG 3.0 (Zimowski, Muraki, Mislevy and Bock, 2003) was used for all of the unidimensional calibrations and for calculating the UIRT(S) *maximum a posteriori* (MAP) scores. *Expected a posteriori* EAP scores were computed for the competency areas under the UIRT(T) method (see Footnote 3). TESTFACT 4.0 (Bock, Gibbons, Schilling, Muraki, Wilson, and Wood, 2003) was used to calibrate the items and produce subscores under the multidimensional (MIRT) model, using the full-information, common factor solution with a varimax rotation.

Comparative Results

The empirical accuracy of the subscores relative to the multivariate normal deviates used to generate the data was not a primary focus, in this study. Instead, this study focused on some practical comparisons of the subscores under each the various computational methods.

Figure 3 shows a dot-density plot of the correlations between the subscores for each competency area and computational method, using a multi-trait, multi-method (MTMM) approach (Cambell and Fiske, 1959) to characterize the correlations. This type of MTMM analysis provides some evidence related to construct validity and trait invariance. Same-trait, different methods (STDM) would be expected to correlate well. There are two clusters of the STDM correlations in Figure 3. One cluster, as expected, is near to 1.0. The other (smaller) cluster is located nearer to –0.50. That second anomalous cluster represents the correlations between the ZX, UIRT(T), and UIRT(S) subscores and the MIRT subscores for Area #3. The correlations between the ZX, UIRT(T), and UIRT(S) subscores, alone, are near to 1.0. The anomaly was isolated for that one cluster of subscores that included the MIRT subscore in Area #3. Nonetheless, it highlights the complexity of comparing MIRT subscores[4].

The different-trait, same-method (DTSM) correlations plotted in Figure 3 show one tight cluster in the range 0.35 to 0.45. This cluster represents the

[4] The factor scores from TESTFACT had to be interpreted in terms of the STDM correlations to properly align them with the appropriate competency areas. In practice, MIRT factors can be misinterpreted if the items are not specifically constrained to load on particular factors. Unfortunately, TESTFACT does not provide the capability to implement such constraints.

attenuated correlations between observed subscores measuring the four different competencies. If we were to dissattenuate those correlations, they would closely approximate the 0.50 between-factor correlations used to generate the data. There are also some isolated DTSM correlations nearer to zero. These largely represent the anomaly noted above with respect to the MIRT subscores for Area #3.

The last set of DTDM correlations represent the different-trait, different-method associations. These correlations typically provide base line information for interpreting the other correlations. That is, there is no reason to expect large correlations between the subscores computed for different competency areas using different methods. In this study, the DTDM were reasonable—falling with the range, 0.0 to 0.50.

Another criteria for evaluating the various scoring methods relates to the measurement errors. Table 2 presents the average standard errors for the four competency subscores under each of the four computational methods. The UIRT(T) subscores have the largest standard errors, across the competencies. The measurement errors associated with the standardized number-correct scores, Z(X), are somewhat smaller than the average UIRT(T) errors, however, the are still somewhat large relative to the other two methods. With the minor exception of Area #1, the UIRT(S) scores appear to provide the most accurate estimates. The MIRT results are likewise reasonable, but hardly seem to justify the added complexity of using a multidimensional model.

It was also informative to review diagnostic score profiles for select candidates. Figure 4 provides multivariate score profiles for four simulated candidates. Scores are plotted on the ordinate axis. The four competency areas are represented along the abscissa. The two plots on the left side of the display show two examinees with exactly the same total-test number-correct scores (X=32). The two right-side plots show two other candidates with total-test scores of 46. The upper plots indicate that those two examinees had more homogeneous. The lower plots show far more heterogeneous subscore patterns.

Since the multidimensional generating parameters where known, it was possible to actually compare the more heterogeneous profiles to "truth". In both cases, the two unidimensional methods, UIRT(T) and UIRT(S), produced diagnostic subscore profiles that were more consistent with the true profiles. These profile plots make one thing quite clear – the choice of scoring method can make a rather substantial difference as to what remedial recommendations are communicated to the examinees when performance is relatively heterogeneous across the subscores.

Obviously, these comparisons from a single simulation study, using model-generated data, are not conclusive. However, the results do serve to highlight some of the complexities related to choosing a scoring model for multidimensional subscore reporting. First, have a clear definition of the competencies. "Content dimensions" are only relevant if scores are to be reported strictly in terms of content-defined competencies. Also realize that the

competency dimensions, themselves, can be composite traits. Second, a thorough evaluation of the test information functions for the item banks, on each test form, and for each competency area can help set practical constraints on whether or not reliable subscores can even be computed for some competency areas. Third, it is important to evaluate competing scoring model in terms of the underlying competency structure that will be used to report scores. Fit is one criterion (e.g., conducting a confirmatory factor analysis), but a multi-trait, multi-method analysis and a comparison of conditional standard errors [on a common scale] may uncover anomalies with one or more scoring models. Fourth, simplicity is good. Unidimensional models tend to be easier to manage than multidimensional models and have far more software resources available to aid in calibrating and equating. At the same time, basing all scores on a total-test calibration is not necessarily the best way to go. Maintaining separate metrics for each of the competency traits can have advantages, as demonstrated here. Finally, it is important to evaluate as many individual score profiles as possible, looking for consistency (or lack there of) in how the profiles would be interpreted by examinees or other constituents. In addition to using real data, simulation data created under one or more viable multidimensional generating functions can provide interesting comparative results. A laudable goal is invariance of the scoring model. Cronbach and Gleser (1953) provide an excellent discussion of statistical techniques for evaluating the similarities between multivariate profiles.

Ultimately, the decision to use a given method to compute diagnostic score should blend technical sophistication with operational needs (e.g., software availability). In some cases, where the reliability or validity of scores are questionable, the best choice may be to not report scores.

Presenting Diagnostic Score Profiles for Credentialing Examinations

There is a surprising lack of empirical literature on graphical and tabular presentation techniques (Wainer and Thissen, 1981;Wainer, 1997), despite the proliferation of powerful statistical and graphics packages that offer a wide range of methods for graphing multivariate data. In diagnostic score reporting for credentialing examinations, the literature is almost nonexistent.

Producing graphics amounts to creating *abstract* meaning from quantitative data using aesthetics and mathematics (Tufte, 1997). There are three perceptual aspects involved when interpreting graphs. *Detection* deals with basic information from the data that must be discernable in the graph. Facilitating detection means maximizing the unique and important information, minimizing noise or extraneous information, and eliminating redundancy. *Assembly* is the process of discerning patterned regularities and relationships among the discrete graphics elements. Graphics that facilitate assembly must convey the optimal amount of variability to show the required patterns, avoid eliminating meaningful variation (i.e., blips), avoid displaying extraneous variation or patterns from nothing (e.g., inappropriate use of smoothers or fitting functions),

and avoid information overload. Finally, *perceptual estimation* involves comparing and contrasting the relative magnitudes of two or more elements contained in a graph. Human perceptions can be biased and distort the information. *Scales are important!!!* Common scales are needed for comparisons. Geometric and aesthetic attributes also matter. Lines are usually more accurately perceived than areas or volumes. Color and shading may confuse (e.g. perceptions of red as "hot"). There are a number of excellent resources on perceptual components of graphing and methods for producing good graphics, including Cleveland, (1994), Jacoby (1997, 1998), Tufte (1990, 1997, 2001), Tukey, (1977), Wainer, (1997), Wainer, Hambleton, and Meara, (1999), and Wainer and Thissen (1981).

Summarizing across a number of sources, there are [at least] eight basic rules for good graphics.

1. Show the data or legitimate patterns that represent the data.
2. Induce viewer to think about the substance of the information, not technique or technology.
3. Avoid distortions
4. Codify and make coherent large data sets.
5. Encourage visual comparisons.
6. Reveal the data in multiple layers of detail.
7. Clearly serve a purpose to describe, explore, tabulate, or decorate.

8. Make sure that the graphic(s) is/are closely integrated with statistical and verbal descriptions of results.

When reporting diagnostic subscores, it is essential to keep the purpose and the population of viewers at the forefront during the report design. The tendency in diagnostic score reporting is to underreport useful information. For example, a score report may cover up useful variation or overly simplify a profile of scores. Complicated graphs can be informative, if properly explained. At the same time, perceptions can be altered , appropriately or not, through cuing and mechanisms that direct attention to specific information. Ultimately, providing diagnostic score feedback to candidates on a credentialing examination should help them comprehend, in an unambiguous way, their strengths and weaknesses in terms of the relevant competencies. Comprehension or lack of comprehension by the candidates of the intended message(s) of a score is seldom tested and too often, assumed.

There are three questions that seem relevant to examinees AFTER taking a credentialing examination with moderate to high stakes. First, did I pass or fail? Second, if I failed, how badly did I do? Third, what do I need to study most or practice the most in order to pass when I retake the examination? There is also some subtle psychometric information that testing organizations need to convey to the candidates about their scores (e.g., that there is uncertainty in every score). Ultimately, the score report should provide the relevant feedback in a concise

and meaningful way to answer the examinee's anticipated questions. Some general guidelines for profiling diagnostic information is as follows.

1. Standardize the individual scales and use a reasonable and metric for all of the scores (e.g., a mean of 50 and standard deviation of 10). There are strong arguments often made in favor of a "less is more" philosophy about the number of possible score points. For example a scale of 0 to 20 may be better than a scale of 0 to 100, especially for short subtests.
2. Be careful showing the relationship of subscores to the cut score. Candidates may wish to know how badly they did on particular sections, relative to the cut score. That type of information may be misleading, since the total-test metric is substantively and statistically different (except in the case of the UIRT(T) scoring method) than the subscore metrics. If the cut score is provided in conjunction with the diagnostic subscores, thoroughly audit the profiles of candidates to ensure that some individually aberrant profiles for failing examinees do not might convey the faulty impression that the candidates had more "passing" subscores than "failing" subscores.
3. Add measurement error bars that accurately portray the reliability of every score. Use conditionally measurement errors, if possible.

4. Consider using percentile ranks to report total scores. They can accurately indicate performance relative to the examinee population. The cut score can likewise be shown for the population[5]

5. Consider ordering the variables in terms of either "confidence" (error variances) or magnitude of improvement needed. This approach can be applied to graphics as well as tables.

Obviously, there are operational considerations in terms of systems and software for generating diagnostic score reports. However, with current statistical and graphics computing and printing technologies, there should not serious limitations voiced by competent systems designers. That means that test developers have a number of techniques from which to choose: score tables, profile plots, and narrative text. Narrative test is not addressed here.

Using Tables

Tables are generally good for rank order comparisons (best and worst). As Wainer and Thissen (1981) noted, the rows should be ordered with respect to some aspect of the data (e.g., the magnitude of scores or the magnitude of measurement errors), numbers should be generously rounded, and row spacing or other text manipulations used to "chunk" or highlight relevant sections of the table. Table 3 shows two side-by-side tables displaying the scores for the examinee plotted in the lower right corner of Figure 3. The scores and standard

[5] There is no empirical evidence, that I am aware of, that suggests that providing percentile ranks, in place of or in addition to total test scores, confounds the notion of a "content-based" or "absolute" standard of competency. Percentile ranks have a long history as being useful information to candidates. Their use in credentialing test settings seems appropriate.

errors in the table were arbitrarily scaled to a population mean of 50 and standard deviation of 10. The left-hand table presents the data in competency area order. The right side table presents the same data, but sorted in descending order by the values scale score. Strengths are therefore listed first, followed by weaknesses. The first score is furthermore bold-faced to (possibly) indicate performance that is statistically higher than then population mean.

Using Multivariate Profile Plots

Graphics that display profile plots of the diagnostic scores are very easy to produce and generally interpretable by most examinees. Figure 4 illustrates a simple multivariate score profile for a low performing examinee with number-correct scores of 10, 9, 5, and 8, respectively, on the four competency areas. The scores plotted are based on UIRT(S) scoring and have been further scaled to an arbitrary population mean of 50 and standard deviation of 10. Error bands[6] have also been added about the subscores to reflect the measurement errors associated with each score. Unfortunately, Figure 4 conveys no comparative information about the distribution of scores in the examinee population.

Figure 5 provides enhanced score report information. The right side portrays the same score profile as Figure 4, but uses a different style of bar to de-emphasize the score and, instead, emphasize the possible range of scores within the error band. Horizontal grid lines have been added, including a solid grid

---

[6] Although sometimes complicating the production of tables and graphics, conditional errors of measurement should be used where possible to appropriately reflect the error variance across the subscore scales.

line demarcating the population means for each of the diagnostic scores. The percentile table at the right shows the examinees approximate rank in the population and the approximate location of the cut score. The primary intent is to show how far the examinee is away from the population-based cut score. Spacing enhancements and standard errors for the percentile could likewise be added (see, for example, suggestions by Wainer and Thissen, 1981 and Wainer, Hambleton, and Meara, 1999). Is this an "optimal" score report. Of course not. However, it does convey multiple layers of information at appropriate degrees of detail to answer most of the candidate questions raised earlier.

Concluding Comments

Technological innovations in statistical computing, graphics, and printing make it almost impossible to seriously argue that producing high quality score reports is not feasible, even for relatively small testing programs. Today, a notebook computer, a good statistical graphics package, access to a fast inkjet or laser printer, and a statistical report designer with reasonable database management and programming skills can generate fairly high volume, publication quality score reports.

I have tried to convey the message that "high quality" is not about flashy images. High quality score reporting is a merger of sound computational methods that produce reliable and valid data (the subscores) with appropriate graphical design to generate tables and graphics that convey intentional

information and avoid distortion and misinterpretation. Credentialing examinations have some well-known limitations related to diagnostic reporting score, such as item banks that are designed strictly to maximize the precision of mastery decisions. Yet, despite the practical limitations in certification and licensure testing, it should be possible to find an appropriate diagnostic scoring models through experimentation and simulation. When combined with informative graphics and tables, there is no reason to limit the information provided to candidates to a cryptic message like, "Sorry, you fail!"

References

Author. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: Jointly published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

Bock, R. D, & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46,* 443-458.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0.* [Computer Program]. Chicago, IL: Scientific Software International.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cleveland, W. W. (1994). *The Elements of Graphing Data, 2nd Ed.*. Summit, NJ: Hobart Press.

Cronbach, L. J. & Gleser, G. C. (1953). Assessing similarity between score profiles. *Psychological Bulletin, 50(6)*, 456-473.

Gibbons, R. D. & Hedecker, D. R. (1992). Full-information item bi-factor analysis. Psychometrika, 57, 423-436.

Hambleton, R. K & Swaminathan, H. (1985). *Item Response Theory*. Boston, MA: Kluwer.

Jacoby, W. G. (1997). *Statistical Graphics for Univariate and Bivariate Data*. Thousand Oaks, CA: Sage Publications. (Monograph #117)

Jacoby, W. G. (1998). *Statistical Graphics for Visualizing Multivariate Data*. Thousand Oaks, CA: Sage Publications. (Monograph #120)

Luecht, R. M. (1996). Multidimensional Computer-Adaptive Testing in Certification and Licensure Settings. *Applied Psychological Measurement*.

Luecht, R. M. & Nungester, R. J. (1998). Some practical applications of computerized adaptive sequential testing. *Journal of Educational Measurement*, 35, 229-249.

Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago, IL: University of Chicago Press.

Tufte, E. R. (1990). *Envisioning Information: Narratives of Space and Time*. Cheshire, Conn: Graphics Press.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence, and Narrative*. Cheshire, Conn: Graphics Press.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information, 2nd Edition*. Cheshire, Conn: Graphics Press. (Ordered by the UNCG Bookstore)

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wainer, H. (1997). *Visual Revelations*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: a redesign and validity study. *Journal of Educational Measurement, 36*, 301-335.

Wainer, H. & Thissen, D. *Graphical Data Analysis*. Annual Review of Psychology, 32, 191-241.

Xing, D., & Hambleton, R. K. (2001, April). *Impact of several computer-based testing variables on the psychometric properties of credentialing exams*. Paper presented at the meeting of NCME, Seattle.

Zimowski, M. F, Muraki, E., Mislevy, R. J. & Bock, R. D. (2003). *BILOG-MG 3*. [Computer Program]. Chicago, IL: Scientific Software International.

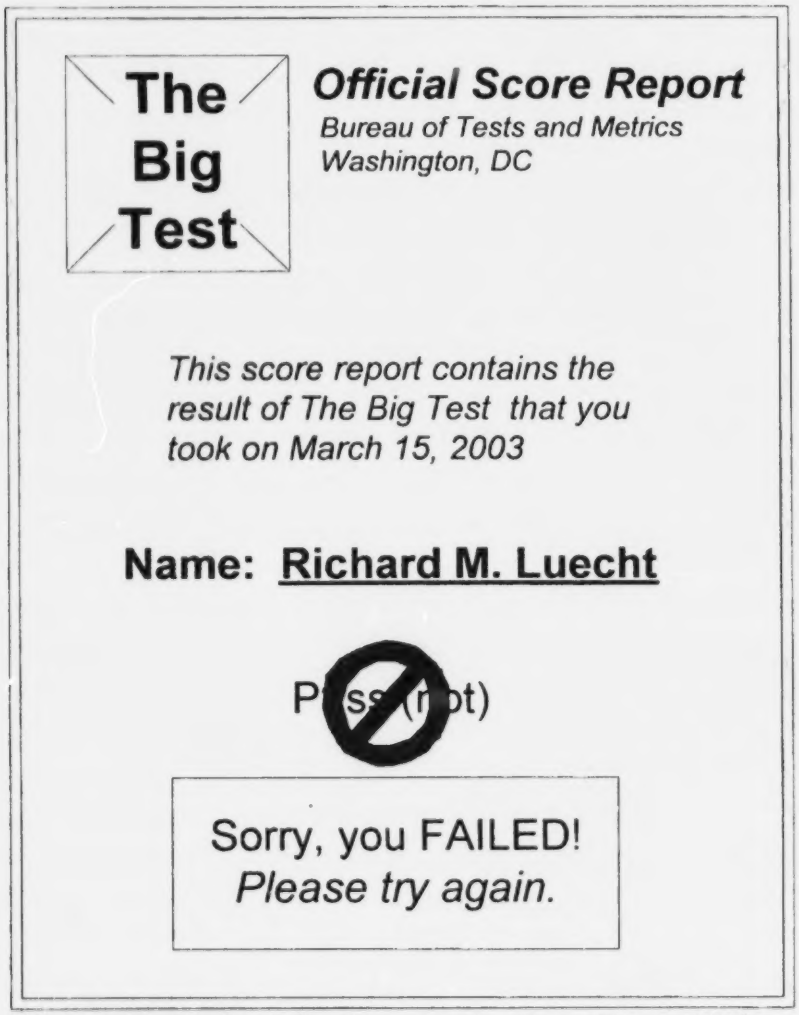
The Big Test
Official Score Report
Bureau of Tests and Metrics
Washington, DC
This score report contains the result of The Big Test that you took on March 15, 2003
Name: Richard M. Luecht
Pass (not)
Sorry, you FAILED!
Please try again.

Figure 1. A Credentialing Examination Score Report Without Scores

BEST COPY AVAILABLE

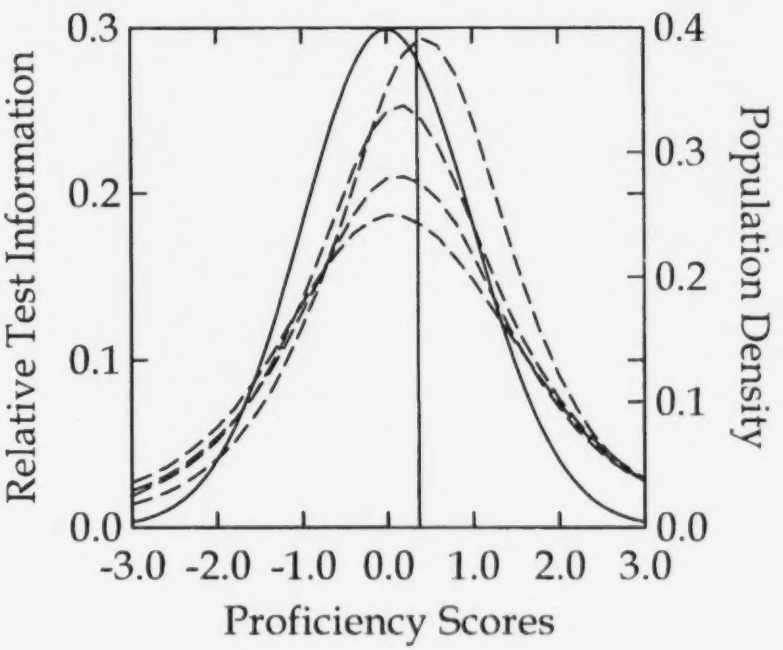

Figure 2. Relative Test Information Curves for Four Content Areas (Dashed Curves) and the Distribution of Proficiency for a Normal Population (Solid Curve)

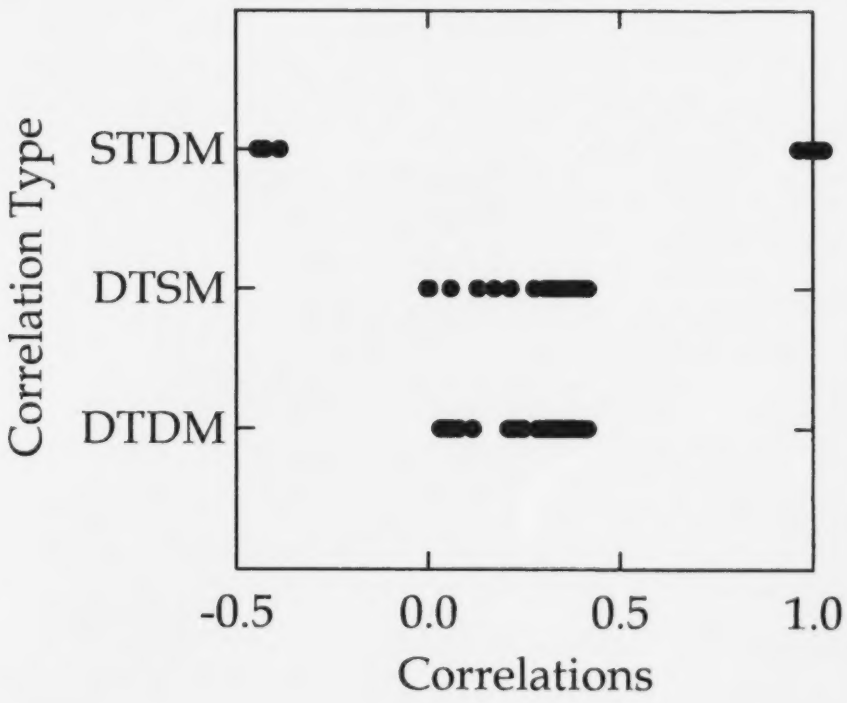

Figure 3. Multitrait-Multimethod Correlations Among the Subscores (STDM=Same Trait, Different Method, DTSM=Different Trait, Same Method, DTDM=Different Trait, Different Method)

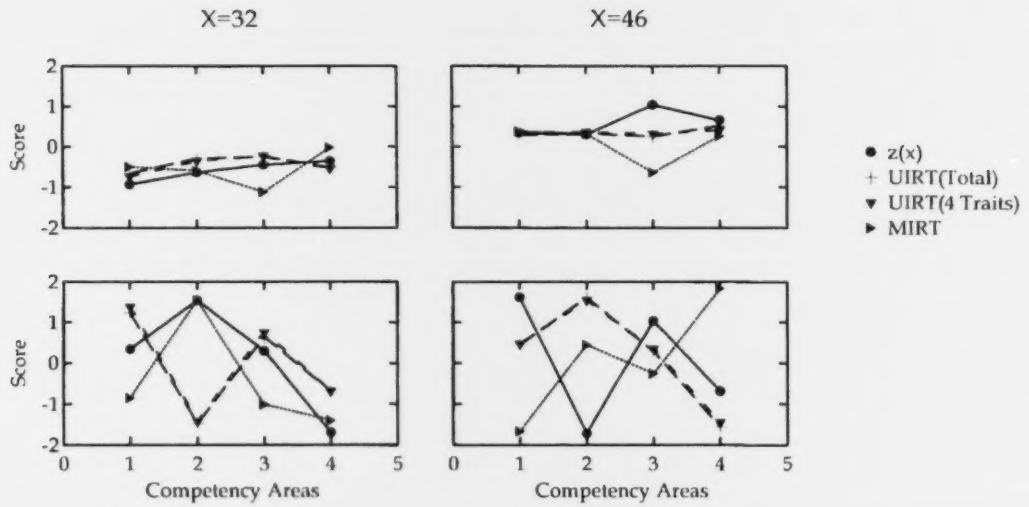

Figure 3. Multidimensional Subscore Profiles for Four Scoring Methods

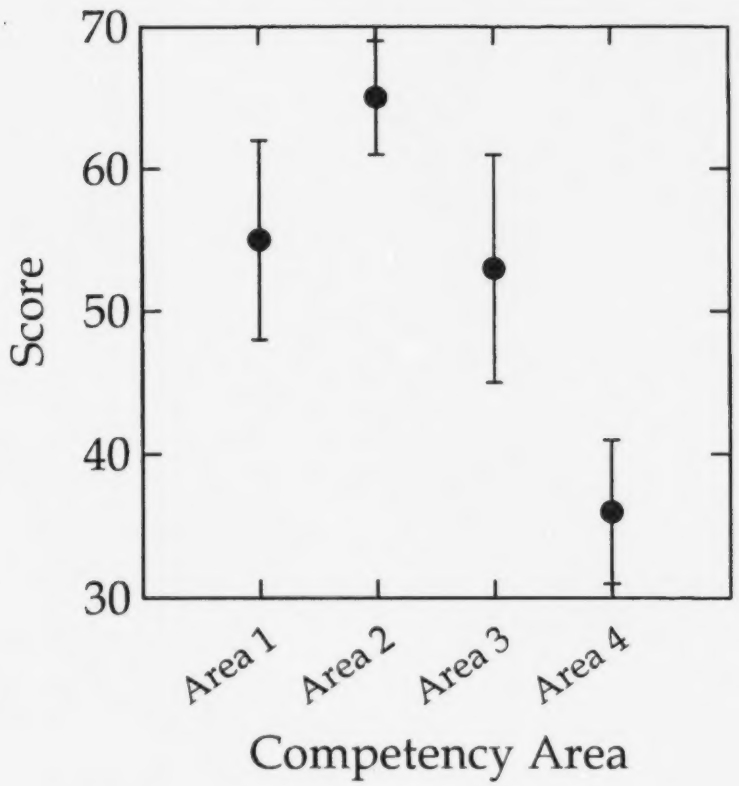

Figure 4. A Profile Plot with Error Bands

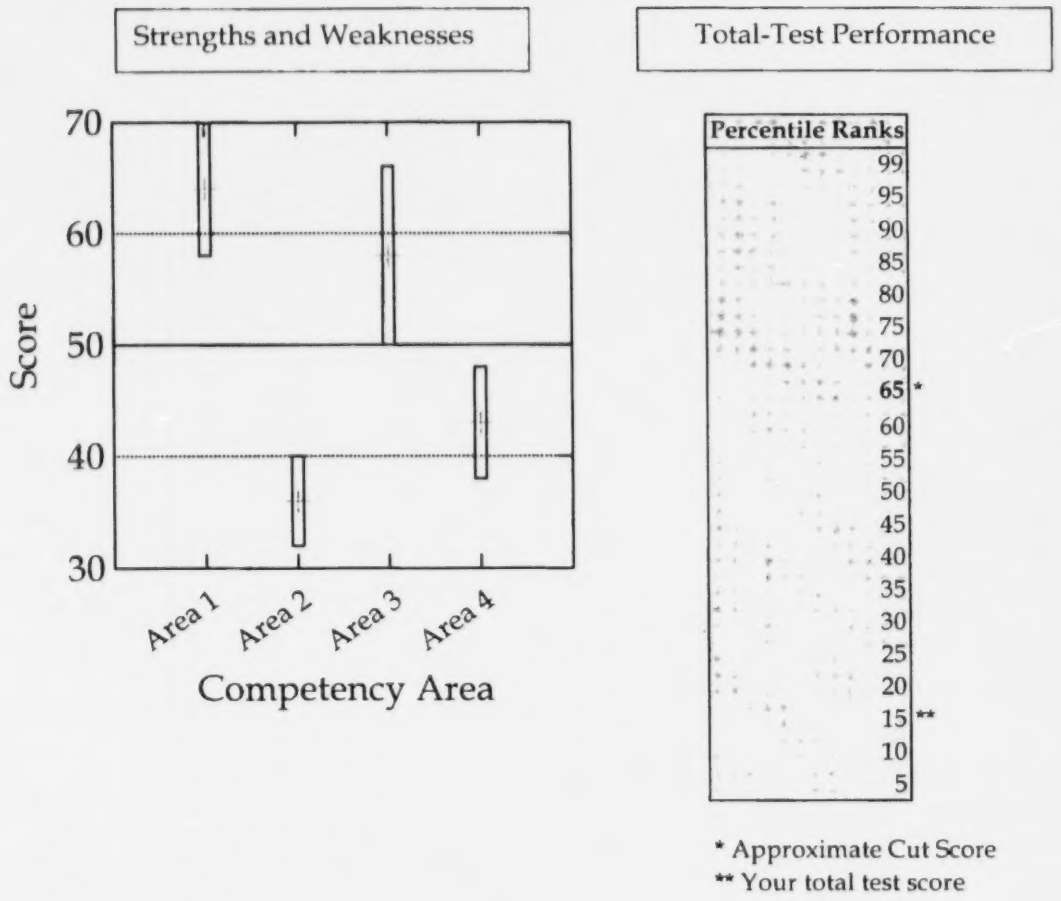

Figure 5. A Single-Panel Display of Total Test Performance and Diagnostic Information

BEST COPY AVAILABLE

Table 1. Descriptive Statistics for 74 Items, Four Competency Areas: #1 to #4

Areas	*Statistic*	*a*	*b*	*c*
1	N of cases	14	14	14
	Minimum	0.18	-4.38	0.04
	Maximum	1.04	0.49	0.28
	Mean	0.55	-0.80	0.14
	Standard Dev	0.27	1.44	0.08
2	N of cases	38	38	38
	Minimum	0.23	-2.99	0.02
	Maximum	1.37	1.07	0.41
	Mean	0.66	-0.23	0.15
	Standard Dev	0.24	0.85	0.11
3	N of cases	7	7	7
	Minimum	0.25	-3.01	0.05
	Maximum	0.89	1.06	0.38
	Mean	0.65	-0.50	0.21
	Standard Dev	0.22	1.35	0.14
4	N of cases	15	15	15
	Minimum	0.18	-0.65	0.02
	Maximum	1.17	1.93	0.36
	Mean	0.75	0.33	0.19
	Standard Dev	0.27	0.64	0.11

Table 2. Average Standard Errors (Adjusted[1])

	Competency Areas			
Method	1	2	3	4
Z(X)[2]	0.67	0.42	0.78	0.54
UIRT(T)	1.06	0.43	1.61	0.76
UIRT(S)	0.39	0.15	0.57	0.28
MIRT	0.20	0.34	0.66	0.38

[1] Adjusted for scale differences. Values shown are based on unit normal distributions

[2] Standard errors based on unconditional reliability (Cronbach's α)

Table 3. Comparison of Score Tables (Left Side in Competency Area Order, Right Side in Score Order)

Competency Area	Score	Error
Area 1	55	7
Area 2	65	4
Area 3	53	8
Area 4	36	5

Competency Area	Score	Error
Area 2	**65**	**4**
Area 1	55	7
Area 3	53	8
Area 4	36	5

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM034853

I. DOCUMENT IDENTIFICATION:

Title: Applications of Multidimensional Diagnostic Scoring for Certification and Licensure Tests	
Author(s): Richard M. Luecht	
Corporate Source: University of North Carolina at Greensboro	Publication Date: April 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Richard M Luecht	Printed Name/Position/Title: Richard M. Luecht, Professor	
Organization/Address: Department of Educational Research Methodology, UNCG, 209 Curry Building, Greensboro NC 27402-6170	Telephone: 336/334-3473	Fax: 256-0405
	E-Mail Address: rmluecht@uncg.edu	Date: 4/28/03

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
University of Maryland, College Park
1129 Shriver Lab
College Park, MD 20742